Medical Analysis
Fall 2023
Updated 9/12/2023

Background

Using the HeartFailure data, determine the factors that influence heart failure.

Dataset

Use the dataset HeartFailure spreadsheet. This data is from the University of California Irvine machine learning dataset repository. The dataset is on Blackboard.

https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

Data Fields

Age	age of the patient (years)
Anaemia	decrease of red blood cells or hemoglobin (boolean)
High blood pressure	if the patient has hypertension (boolean)
Creatinine phosphokinase (CPK):	level of the CPK enzyme in the blood (mcg/L)
Diabetes	if the patient has diabetes (boolean)
Ejection fraction	percentage of blood leaving the heart at each contraction (percentage)
Platelets	platelets in the blood (kiloplatelets/mL)
Sex	woman or man (binary)
Serum creatinine	level of serum creatinine in the blood (mg/dL)
Serum sodium	level of serum sodium in the blood (mEq/L)
Smoking	if the patient smokes or not (boolean)
Time	follow-up period (days)
Death event	if the patient deceased during the follow-up period (boolean)

Data Values

Sex - Gender of patient Male = 1, Female =0
Age - Age of patient
Diabetes - 0 = No, 1 = Yes
Anaemia - 0 = No, 1 = Yes
High_blood_pressure - 0 = No, 1 = Yes
Smoking - 0 = No, 1 = Yes
DEATH_EVENT - 0 = No, 1 = Yes

Assignment

What's due:

PowerPoint presentation due before class on Monday, October 2, 2023. The expected length of presentation is 12-15 minutes, approximately 20 slides. Please send me the slides at least one hour before class. You can describe the slides from your seat and use a remote control to advance the slides.

Outline

Using the medical dataset, perform a thorough analysis of the following aspects of the data.

1. Visualization (25 points)
   a. Develop an overall view of relationship of dependent variable (Death event) with all continuous variables (see iris example with scatterplot matrix in Rcmdr) (5 points)
   b. Highlight at least two graphs where there are strong relationships between pairs of continuous variables (see iris example with scatterplots in Rcmdr) (10 points)
   c. Show at least two graphs with pairs of continuous variables and plot Death events (see iris example in RStudio) (10 points)
2. Perceptrons (25 points)
   a. Develop at least two perceptrons where there are two x-variables, using continuous and binary x-variables, with Death event as the y-variable; calculate a measure of accuracy for each (10 points)
   b. Develop at least two perceptrons where there are three or more x-variables and Death event is the y-variable; calculate a measure of accuracy for each (10 points)
   c. Summarize the accuracy of all the perceptron models (5 points)
3. Support Vector Machines (SVM) (35 points)
   a. Develop at least two SVMs where there are two x-variables, using continuous and binary x-variables, with Death event as the y-variable; calculate a measure of accuracy for each (10 points)
   b. Develop at least two SVMs where there are three or more x-variables, with Death event as the y-variable; calculate a measure of accuracy for each (10 points)
   c. Generate a graphic showing the results of each of the SVM results (10 points)
   d. Summarize the accuracy of all the SVM models (5 points)
4. Identify a list of lessons learned (15 points)
   a. When do visualizations help in this specific example? When do they not help in this specific example? (5 points)
   b. Which choice model techniques (perceptrons and SVM) worked in this specific example? Which did not? Why? (5 points)
   c. What other techniques could you use (other than perceptrons and SVM)? (5 points)